16
Looking at Elementary Analyses

		Exh	Page
Introduction 16-1			
		A	
		1	16-4
16A. Four thermometers 16-3		2	16-5
		3	16-7
16B. Fitting anyway 16-10		4	16-8
- doing our best 16-11			
		B	
- a more symmetric description 16-11		5	16-12
- a little more data 16-13		6	16-14
		C	
16C. Arithmetic for small tables 16-15		7	16-16/17
- fit by means 16-15		8	16-20
common term 16-18		9	16-22
stripe fits 16-18			
		D	
stripe effect 16-18		10	16-24/25
- easy rough fits 16-19		11	16-27
		12	16-28
16D. Elementary and effect analyses 16-23			
		E	
elementary analysis 16-23		13	16-31
- verbal forms of the alternatives 16-26		14	16-32
- alternative analyses of our 3×4 16-26		15	16-34
- streamlining analysis by means 16-29		16	16-36/37
- parts of fits 16-30			
		F	
		17	16-40
16E. Changing and combining fits; polishing 16-30		18	16-43
- a more general view 16-30			
		G	
- working with medians 16-35		19	16-46
median polish 16-35		20	16-48
- how much polish? 16-38		21	16-49
- other kinds of polish 16-38		22	16-51
- combining analyses 16-39		23	16-53
		24	16-55
16F. Checking two-way analyses 16-39		25	16-57
- the need for checking 16-39			
		H	
- making the check 16-41			
		K	
difference analysis 16-41		26	16-62/63
- correct and go on 16-44		27	16-65/66
		P	
16G. Making two-way plots easier and more effective			
- easier construction 16-45			
- showing outside residuals 16-47			
- summarizing other residuals 16-50			
- looking harder at the pattern 16-52			
**********			
16H. How far have we come? 16-58			
**********			
16K Median polishing without choices (background)			
- full-step polishing 16-61			
- a possible standard 16-64			
**********			
16P. Problems for chapter 16 16-67			

16

Looking at Elementary Analyses

Two-way tables are a little further outside everyday experience than either batches or analyses of y vs. x. We need to become familiar with them from a variety of aspects, including:

- what they are.
- what the results of their simplest analyses look like.
- how to think about these results.
- how to make these analyses.
- how to go further or do better.

This chapter is aimed at the first four of these aspects. It will show you the results of analysis, explain how they can be assembled, and begin to show you how to think about the results. It will tell you something about how to make the analyses, beginning in section 16C.

A two-way table, of the sort we are going to treat here, involves three variables:

- a response.
- two factors.

In our first examples, for instance, the response will be a (mean) monthly temperature and the two factors will be places

on the ground and months of the calendar. For each place, say Washington, and each month, say April, there will be a temperature. In general, for each combination of a version of one factor with a version of the other there will be a response.

The nature of a two-way table, described carefully, is ordinarily just this:

- a man might want to think of the response as caused by the combination of the two factors.

Just writing down a two-way table could not force us to think in terms of cause and effect -- it merely makes it easy for us to do this. In individual cases, we may or may not want to think this way, depending on our knowledge of:

- the factors.
- the response.
- other variables.
- how the two-way table turns out.

We can -- and should -- recognize two-way tables as another step toward mastering data whose structure shows increasing complexity. Exhibit 1 sketches the situation. As it shows, factors began to appear as soon as we started comparing two or more batches, though we did not say so.

The two-way table is the third major key to the analysis of data -- exploratory, confirmatory, or what have you. Not only is it important for its own sake, and for its

exhibit 1 of chapter 16

Stages of Complexity in the Structure of Data

	Variables involved	
Structure	**Number**	**Character**
A batch	1	One response
Several batches	2	One response, one factor (its versions label the batches)
Most y vs x	2	One response, one factor (its versions have been assigned numbers)
Chapter 12 (No distinction among points)	2	Two responses
Chapter 12 (points divided into classes)	3	Two responses, one factor (its versions label the classes
Two-way table of responses	3	One response, two factors

uses in analyzing data in more complex forms, but both thorough understanding and intuitive familiarity are essential steps in the understanding of regression techniques.

Regression techniques, surely our most powerful kit of tools for seeing into data, are not simple to understand or to use with greatest effect. The whole of this account of exploratory data analysis can reasonably be considered an introduction to a later, careful treatment of regression as a tool. The chapters on two-way tables, like those that have gone before and most of those that will follow, are critically important parts of that introducation.

16A. Four thermometers

We are all easily convinced both that it is usually warmer in July than in January and that it is usually warmer in Laredo, Texas than in Caribou, Maine. When we look up the four mean monthly temperatures (at the airports) we find

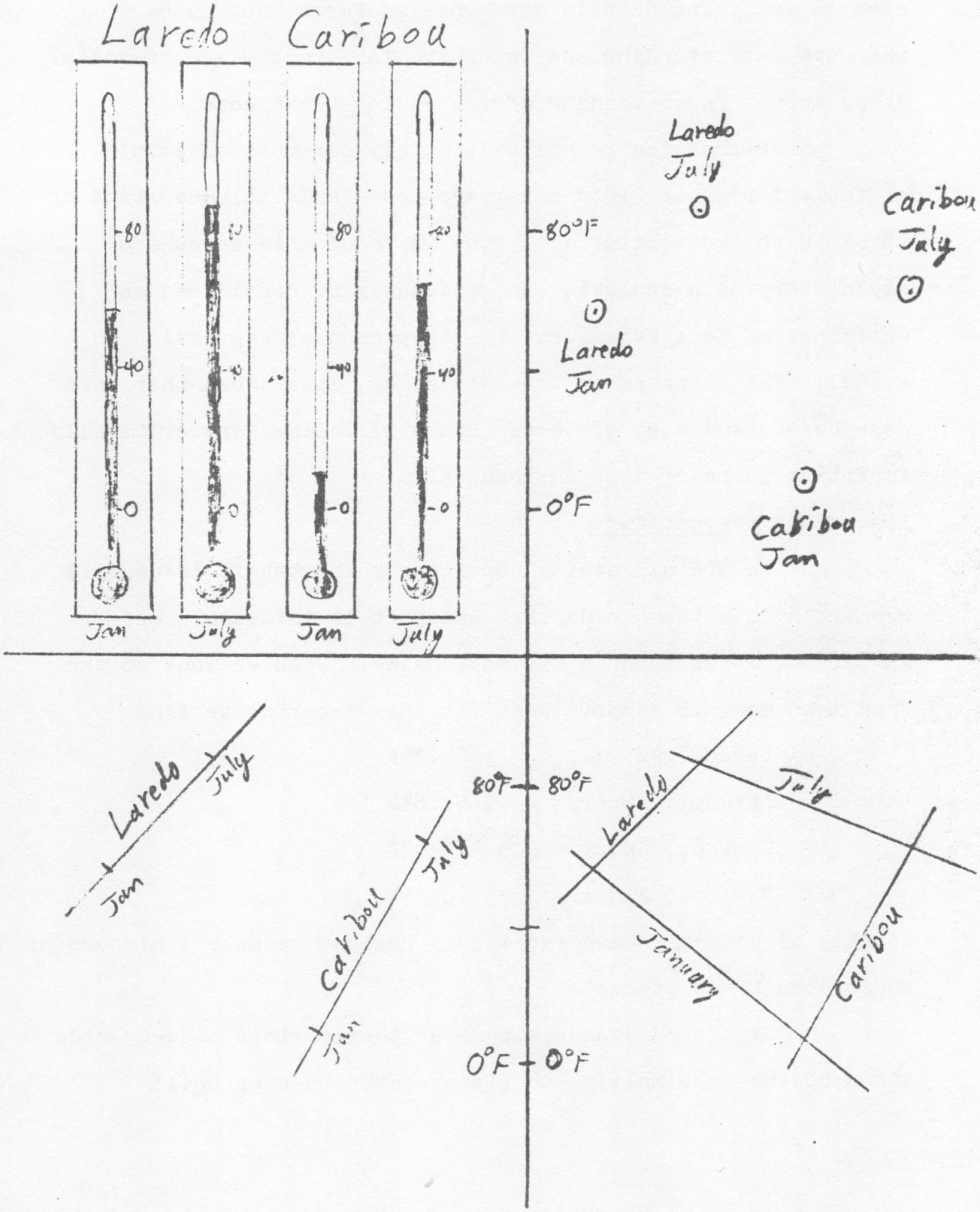
Laredo, July:	87.7°F
Laredo, January:	57.6°F
Caribou, July:	64.0°F
Caribou, January:	8.7°F

roughly as we might have expected. How can we make a picture of these four temperatures?

Exhibit 2 starts with four thermometers side-by-side and proceeds to simplify the picture step-by-step until

exhibit 2 of chapter 16

Four mean monthly temperatures near the East Coast
(from four thermometers to a quadrilateral)



we have four lines intersecting in four points -- one line each for Laredo, Caribou, January, and July, one intersection for each combination of place and month. Right-to-left location on the final picture is "forget it" -- just a residue of how we happened to space the four thermometers. We can change it as we wish, it is purely a matter of convenience. Down-to-up location is meaningful -- it shows temperature. We will soon be used to this kind of picture where one coordinate is meaningful, but the other is only for convenience.

Let us look at another example, one that is simpler in a helpful way. We can stick to January and July temperatures if we go to two airports in Arizona. Yuma is low and hot, in Southwest Arizona. Flagstaff is high and cool, in north central Arizona, just south of the Grand Canyon. The 4 mean monthly temperatures are:

Yuma, January:	55.3°F
Yuma, July:	94.6°F
Flagstaff, January:	25.3°F
Flagstaff, July:	65.2°F

The differences are about as large as for Laredo and Caribou. Exhibit 3 shows, however, that we can make the 4-line, 4-point picture for Yuma and Flagstaff, January and July look very much like a parallelogram. Exhibit 4 shows that once we can make it look like a parallelogram, we can make it look like a rectangle -- by playing with the "forget it" coordinate.

exhibit 3 of chapter 16

Four mean monthly temperatures in Arizona
(to a larger scale, almost a parallelogram)

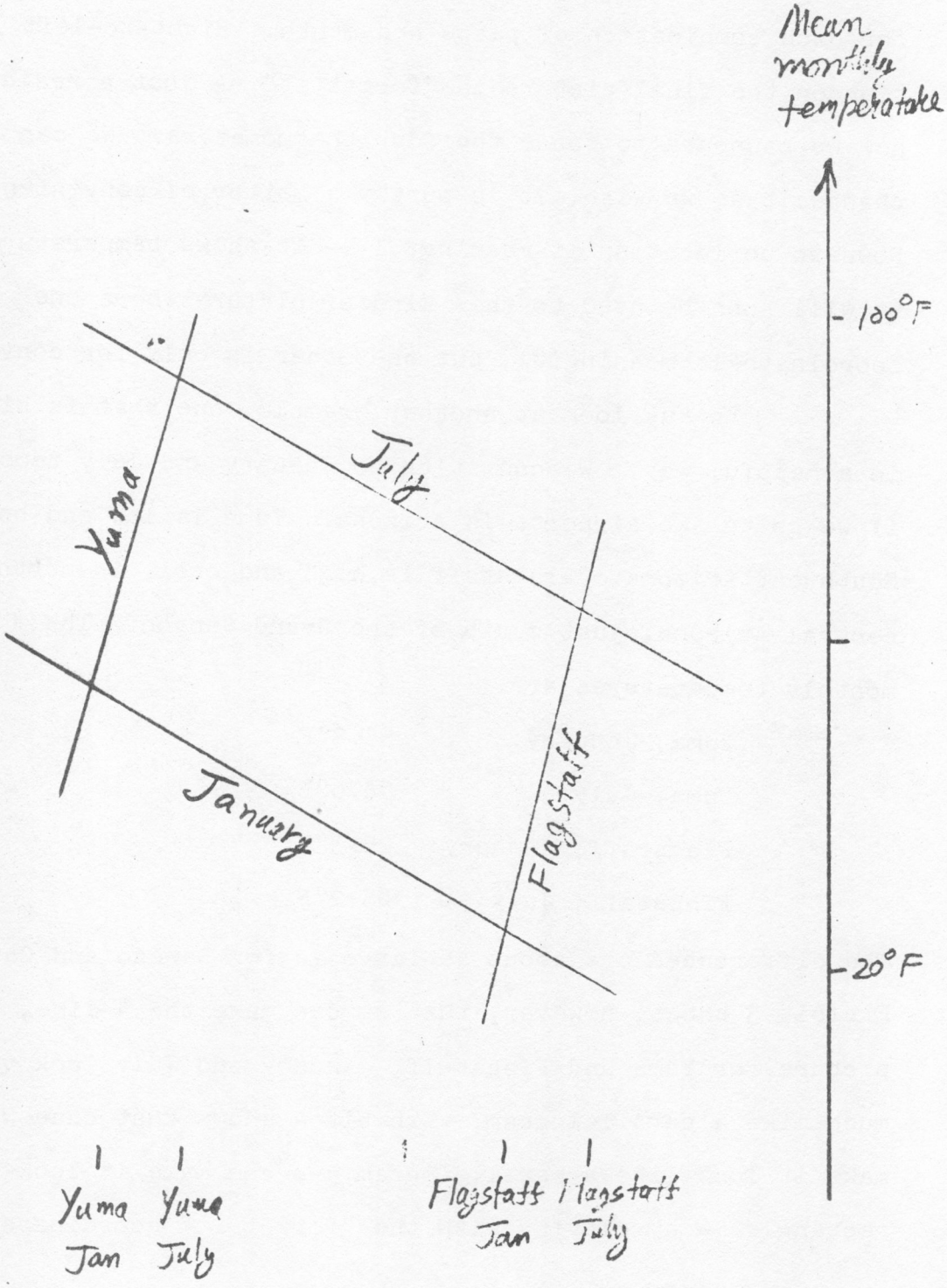
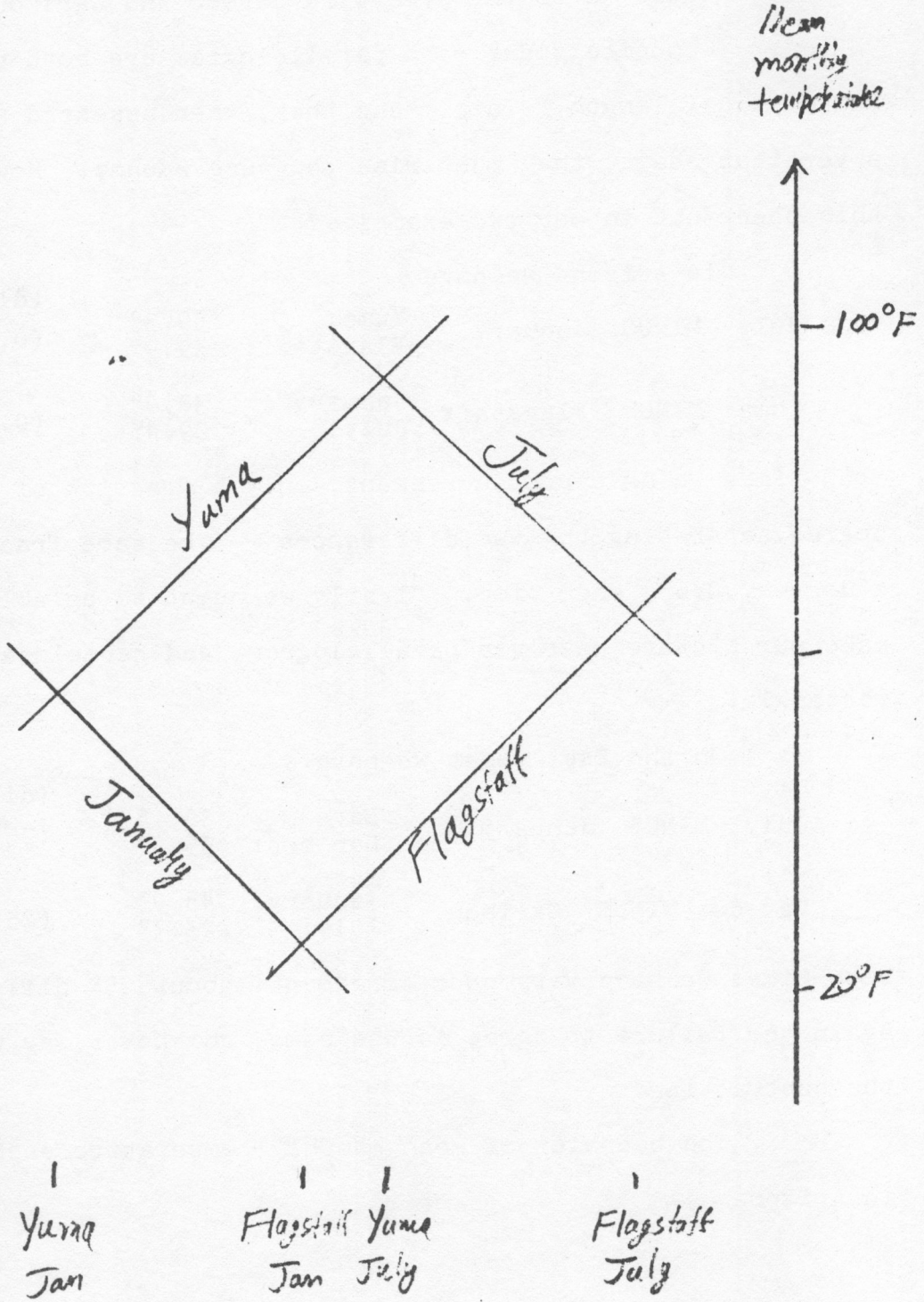


exhibit 4 of chapter 16

The four Arizona temperatures made nearly rectangular



What is it about Arizona temperatures that lets us do this? Could we do the same with Laredo and Caribou?

Opposite sides of a parallelogram are both parallel and of equal length. This means that, when assessed against a vertical scale, they must rise the same amount. How does this check out in our two examples?

In Arizona we have

					(diff)
July	MINUS	January	Yuma:	39.3°	
			Flagstaff:	39.9°	(0.6°)
Yuma	MINUS	Flagstaff	January:	30.0°	
			July:	29.4°	(0.6°)

Both times we have good agreement, only a fraction of a degree separating the two differences -- the same fraction of a degree, 0.6°, each time. Clearly we ought to be able to make our picture nearly a parallelogram, and hence nearly a rectangle.

On the East Coast we have

					(diff)
July	MINUS	January	Laredo:	30.1°	
			Caribou:	55.3°	(25.2°)
Laredo	MINUS	Caribou,	January:	48.9°	
			July:	23.7°	(25.2°)

Both times we have very poor agreement, about 25° difference -- again the failure to agree is the same, whichever way we do the subtraction.

The behavior of mean monthly temperature seems simpler in Arizona.

16B. Fitting anyway

It is simpler to describe -- to a close approximation -- our four Arizona temperatures. Simple in the sense that it seems easy to try to spread the description out to other places and other months.

We can say

- July 39.6°F hotter than January
- Yuma 29.7°F hotter than Flagstaff
- Flagstaff in January 25.4°F.

This describes 25.4°F, 55.1°F, 65.0°F and 94.7°F for the four combinations of place and month. Each of these is within 0.2°F of the values with which we started. As an incomplete description, this is a very close one.

It seems a simple description, too, for we have rather clear anticipations for other Arizona places and for other Arizona months. At another place, we anticipate -- since we know no better -- July to be 39.6°F warmer than January. In another month, we anticipate -- since we know no better -- Yuma to be 29.7°F warmer than Flagstaff.

The simplicity of the parallelogram that may as well be a rectangle is reflected in a meaningful simplicity of expression and thought.

This simplicity comes from breaking the fit into parts.

* doing our best *

What of the East Coast? Here such a simple description cannot do too well. That an incomplete description is far from perfect, however, is no reason for not using it. By making it, we are at least led to write down the residuals that show in what ways and by how much it fails.

When we try to do as well as we can in the same style for Laredo and Caribou, July and January, we find:

- Caribou in January: 15.0°F
- Laredo 36.3°F warmer than Caribou.
- July 24.7°F warmer than January.

As exhibit 5 shows, the results of this incomplete description fail by ±6.3°F to describe the mean temperatures for each of these four combinations of place and month.

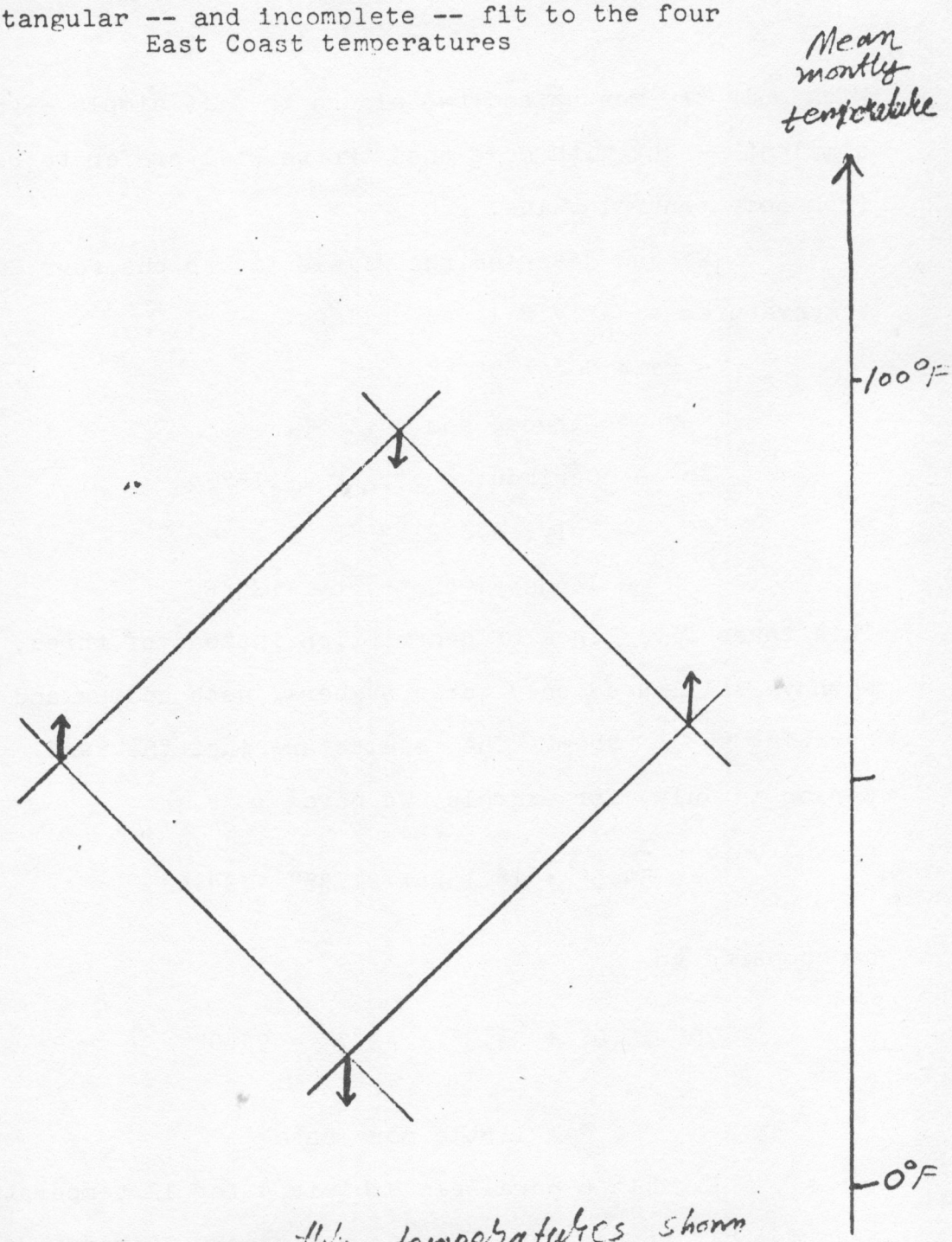
The simple incomplete description is far from perfect. We can -- and in due course will -- do better. But it serves a purpose -- it does give us a general idea of both July-January and Laredo-Caribou differences.

* a more symmetric description *

Let us plan, then, to carry on, using both more places and more months. Before we start, though, we ought to think about the lack of symmetry in the description we have so far used. In each case, we picked the coldest combination of place and month, and reckoned changes upward from there.

exhibit 5 of chapter 16 -- East Coast temperatures

A rectangular -- and incomplete -- fit to the four East Coast temperatures



Actual mean monthly temperatures shown by heights of ends of arrows

With only two months and two places this is simple -- perhaps simplest -- but with more than two we will prefer to reckon from some central value.

We can describe the simple fit to the four East Coast temperatures equally well as follows:

- East Coast 54.5°F.
  - -- Laredo, add 18.15°F.
  - -- Caribou, subtract 18.15°F.
  - -- July, add 21.35°F.
  - -- January, subtract 21.35°F.

This takes five lines of description instead of three, but we have still used only three numbers, both adding and subtracting two of them. The results are just the same. At Laredo in July, for example, we have

$$54.5° + 18.15° + 21.35° = 94.0°$$

as compared to

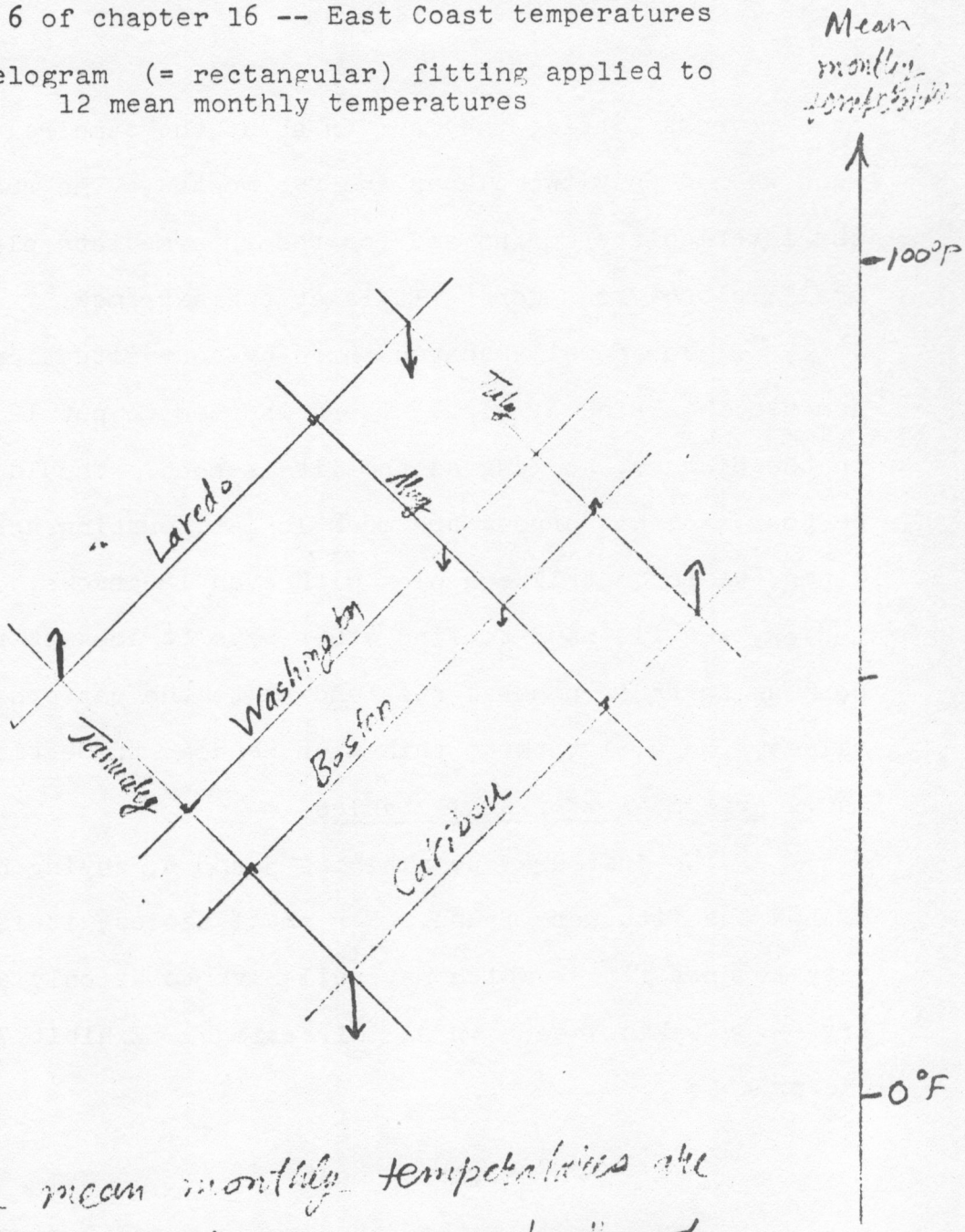
$$15.0° + 36.3° + 42.7° = 94.0°$$

* a little more data *

Exhibit 6 parallels exhibit 4 for 12 temperatures: the mean monthly temperatures for four places -- Laredo, Washington, Boston, and Caribou -- in three months -- January, May, and July. While the lines for Caribou, Laredo, January and July

exhibit 6 of chapter 16 -- East Coast temperatures

Parallelogram (= rectangular) fitting applied to 12 mean monthly temperatures



Actual mean monthly temperatures are shown by the heights of the heads of the arrows.

Fitted mean monthly temperatures are shown by the intersections of lines (= the tails of the arrows)

have moved a little, they are in about the same relation as when we had only two places and two months. The residuals for the intermediate months and for the intermediate places are small; clearly our worst fit is at the extremes.

For a well-behaved three-by-four situation, we can use this kind of plot. There is room to put 12 arrows in the picture, so long as the fit is good. Should the fit be poor, so that arrows are much longer, cutting across lines, we may not like a plot with even 12 arrows. For larger tables, we will need to find other ways to look at residuals, both as to their general size and as to the pattern of their values. We will turn to this when we come to Section 16G.

16C. Arithmetic for small tables

We just drew up exhibits 5 and 6, saying nothing of how the fits were found. For small tables, it is rather easy to do a fit -- which may well have to be only a first try -- by taking means in all directions. Exhibit 7 illustrates the process.

* fit by means *

Panel A gives the data, while panel B gives all the row and column means. Notice that the mean of the column means equals the mean of the row means. This checks the arithmetic, and makes it sensible to call the common value, 57.08 in the example, the grand mean.

exhibit 7 of chapter 16 -- East Coast temperatures

Mean monthly temperatures for four places and three months analyzed by means

A) The DATA -- in °F

	Laredo	Washington	Boston	Caribou
January	57.6	36.2	29.1	8.7
May	81.2	64.7	57.8	48.5
July	87.7	77.3	72.2	64.0

B) MEANS, by PLACE and MONTH -- also in °F

Laredo	75.5	**January**	32.9
Washington	59.4	**May**	63.05
Boston	53.03	**July**	75.3
Caribou	40.4		
(mean)	(57.08)	(mean)	(57.08)

Examples: $75.5 = \frac{1}{3}(57.6+81.2+87.7)$, $32.9 = \frac{1}{4}(57.6+36.2+29.1+8.7)$.

C) MEAN EFFECTS -- means minus grand mean

Laredo	+18.42	**January**	-24.18
Washington	+2.32	**May**	5.97
Boston	-4.05	**July**	18.22
Caribou	-18.68		
(mean)	(0.00)	(mean)	(0.00)

Examples: 18.42 = 75.5-57.08, -24.18 = 32.9-57.08.

D) DESCRIPTION -- rounded to tenths °F

- East Coast 57.1°

  - Laredo, add 18.4°
  - Washington, add 2.3°
  - Boston, subtract 4.0°
  - Caribou, subtract 16.7°

  - January, subtract 24.2°
  - May, add 6.0°
  - July, add 18.2°

(cont'd.)

exhibit 7 of chapter 16 (cont'd.)

E) INCOMPLETE DESCRIPTION, namely FIT

	Laredo	Washington	Boston	Caribou
January	51.3	35.2	28.9	16.2
May	81.5	65.4	59.1	46.4
July	93.7	77.6	71.3	58.6

Examples: 51.3 = 57.1+18.4-24.2, 81.5 = 57.1+18.4+6.0 .

F) RESIDUALS, as given by DATA minus FIT

	Laredo	Washington	Boston	Caribou
January	6.3	1.0	0.2	-7.5
May	-0.3	-0.7	-1.3	2.1
July	-6.0	-0.3	0.9	5.4

G) PROBLEMS

1) Did we need to keep extra decimals in our means? Why or why not?

2) Repeat this analysis with the following set of Arizona mean monthly temperatures:

	Yuma	Phoenix	Prescott	Flagstaff
January	55.3	49.7	35.3	25.3
May	80.4	75.5	60.9	50.0
July	94.6	90.1	76.0	65.2

3) Repeat for the 2×2 table involving Caribou and Laredo, January and July.

So long as we have no more than 3 or 4 numbers to add up at a time -- and nothing worse than 3 or 4 to divide by -- finding means is not too bad. (Even when taking row means, however, it pays to use the technique of chapter 7x. Thus January would involve three sums of one-digit numbers: .6+.2+.1+.7 = 1.6, 7+6+9+8 = 30 and 50+30+20+0 = 100 for a total of 131.6 which halves to 65.8 and halves again to 32.9.)

We will have to deal with the failure of means to be resistant. We can, as we will see, do this by polishing. All in all, then, means are a convenient first start for small tables, up to 3×3, 3×4 and 4×4.

When we have a fit to the table as a whole, a

common term

and values fitted to individual rows or columns

stripe fits

(it is convenient to use "stripe" to mean "row or column"), it is natural to form their differences. Such a difference will be called a

stripe effect

and will satisfy the identity

stripe effect = stripe fit MINUS common term .

In tables we shall almost always write "eff" for effect. We will do this in text occasionally, too.

Panel C converts stripe means into stripe effects, which Panel D states in words. Panel E sets out the values of the fits, as found from:

fit = common term + row effect + column effect,

which is only one of several ways to express the fit. Finally, panel F calculates residuals from

residual = data - fit.

The picture of exhibit 6 came from the arithmetic of exhibit 7.

* easy rough fits *

One way to make fitting by means a little easier is to make every division have an integer answer. This will surely happen if every data value is divisible by the total count, divisible by 12 for a 3×4 table. We can always make an analysis by modifying the data values so that this happens. Exhibit 8 does this for our 3×4. Notice the two-term calculation of the fit from

fit = row effect + column fit

corresponding to

fit = row effect + (column effect + common term) .

In this instance, the loss in quality of the fit due to modifying the data values matters little to us. No residual is altered by more than ±0.5°F. Such changes are likely to be of little importance in the face of four residuals of ±6 or 7°F.

exhibit 8 of chapter 16 -- East Coast temperatures

Quick, rough analysis for our 3×4

A) MODIFIED DATA, and MEANS -- unit 0.1°F

	Laredo	Washington	Boston	Caribou	(mean)
January	576	360	288	84	327
May	816	648	576	480	630
July	876	768	720	636	750
(mean)	756	592	528	400	(569)

Examples: $327 = \frac{1}{4}(576+360+288+84)$; $756 = \frac{1}{3}(576+816+876)$.

B) ASSEMBLY of ROW EFFECTS and COLUMN MEANS into FIT

	756	592	528	400
-242	514	350	286	158
61	817	653	589	461
181	937	773	709	581

Examples: 756 copied; -242 = 327-569; 514 = -242 + 756.

C) RESIDUALS, from ACTUAL DATA (panel A, exhibit 7) -- still in 0.1°F

62	12	5	-71
-5	-6	-11	24
-60	0	-13	59

Examples: 62 = 576-514, -5 = 812-817, -60 = 877-937

Let us be more vigorous. Instead of moving to multiples of 1.2°F (= 12 tenths), let us move to multiples of 12°F. Exhibit 9 has the values and their analysis. The residuals are now not good enough to make us happy. They are, however, quite good enough to begin with. Polishing will cure them, as the result of a later problem will show.

Notice that, for variety, we have made the fit by a different two-term expression, namely

fit = column effect + row fit

corresponding to

fit = column effect + (row effect + common term).

In addition to the two two-term assemblies and the original three-term version, there is one more. While we essentially never use it, symmetry makes it easier to bring in than to leave out. It is

fit = column fit + row fit - common term .

exhibit 9 of chapter 16 -- East Coast temperatures

Quicker, rougher analysis for our 3×4

A) MORE MODIFIED DATA, and its MEANS -- in °F

	Laredo	Washington	Boston	Caribou	(mean)
January	60	36	24	12	33
May	84	60	60	48	63
July	84	72	72	60	72
(mean)	76	56	52	40	56

Examples: $33 = \frac{1}{4}(60 + 36 + 24 + 12)$; $76 = \frac{1}{3}(60 + 84 + 84)$.

B) ASSEMBLY of COLUMN EFFECTS and ROW MEANS into FIT

	20	0	-4	-16
33	53	33	29	17
63	83	63	59	47
72	92	72	68	56

Examples: 20 = 76 - 56; 33 copied; 53 = 33 + 20.

C) RESIDUALS, from ACTUAL DATA (panel A, exhibit 7)

4.6	3.2	.1	-8.3
-1.8	1.7	-1.2	1.5
-4.3	5.3	4.2	8.0

Examples: 4.6 = 57.6 - 53; -1.8 = 81.2 - 83; -4.3 = 87.7 - 92.

16D. Elementary and effect analyses

We are now used to these ideas:

- a fit given in terms of 2 or 3 parts.
- residuals, one for each data value.

By now it should be clear that it is easier to understand the fit in parts than it is to understand the corresponding two-way table of fitted numbers.

We want to have standard forms for writing down residuals and fit. Exhibit 10 shows five ways to do this -- for simplicity using the more modified data of exhibit 8. The first four have already been illustrated or discussed; they are collected here for comparison. The fifth combines all the first four in one convenient package, offering us both fit and effect for each stripe. It is an

elementary analysis

an example of the form we shall most often use. When we give less detail, an effect analysis, as in panel A, will be most usual.

An elementary analysis "has everything". One way to describe its properties is this:

> Any four numbers at the corners of a rectangle, one corner among the residuals and another of the four lower-right values, sum to the data value corresponding to that residual. Accordingly, all three corners except the residual sum to the fit.

exhibit 10 of chapter 16 -- East Coast temperatures

Five ways to write down the same fit and residuals
(Data of panel A of exhibit 9)

A) ROW and COLUMN EFFECTS and COMMON TERM -- an EFFECT ANALYSIS

	Laredo	Washington	Boston	Caribou	eff
January	7	3	5	-5	-23
May	1	-3	1	1	7
July	-8	0	4	4	16
eff	20	0	-4	-16	56

Examples of fit: 53 = (-23) + 20 + 56
83 = 7 + 20 + 56

B) ROW EFFECTS and COLUMN FITS

	Laredo	Washington	Boston	Caribou	eff
January	7	3	5	-5	-23
May	1	-3	1	1	7
July	-8	0	4	4	16
fit	76	56	52	40	(00)

Examples of fit: 53 = (-23) + 76; 83 = 7 + 76

C) ROW FITS and COLUMN EFFECTS

	Laredo	Washington	Boston	Caribou	fit
January	7	3	5	-5	33
May	1	-3	1	1	63
July	-8	0	4	4	72
eff	20	0	-4	-16	(00)

Examples of fit: 53 = 33 + 20; 83 = 63 + 20 .

(cont'd.)

exhibit 10 (cont'd.)

D) ROW and COLUMN FITS, minus COMMON TERM

	Laredo	Washington	Boston	Caribou	fit
January	7	3	5	-5	33
May	1	-3	1	1	63
July	-8	0	4	4	72
fit	76	56	52	40	(-56)

Examples of fit: 53 = 33 + 76 + (-56);
83 = 63 + 76 + (-56)

E) ALL COMBINATIONS -- an ELEMENTARY ANALYSIS

	Laredo	Washington	Boston	Caribou	eff	fit
January	7	3	5	-5	-23	33
May	1	-3	1	1	7	63
July	-8	0	4	4	16	72
eff	20	0	-4	-16	56	00
fit	76	56	52	40	00	-56

Examples of fit: Given in panels above.

* verbal forms of the alternatives *

Each of the ways to assemble a fit corresponds to a different way of describing that fit in words. Exhibit 11 gives descriptions in words for each of the four different forms of assembly illustrated in exhibit 9.

If a fit can be described in any one of these ways, it can be described in all four, with appropriate changes in numbers.

It pays to go over these alternative verbal descriptions enough times to begin to have a feeling for what this sort of fit means and behaves like.

* alternative analyses of our 3×4 *

Let us return to the detailed data. The rougher approximations have made it easy for us to write down simple examples. It has certainly paid its way. Now they can serve us in another way -- illustrating why we say "an" elementary analysis (or effect analysis) and not "the" elementary analysis.

Exhibit 12 sets out 3 different effect analyses -- and then the 3 corresponding elementary analyses -- for our 3×4 example. Within each set of three, the analyses are clearly different. Yet they are analyses of exactly the same data.

When a fit was just a single number -- a central value for a batch -- we clung to

data = fit + residual

but we admitted that there could be different kinds of fits.

exhibit 11 of chapter 16 -- East Coast temperatures

Verbal descriptions corresponding to the four different forms of assembly of the same fit (analyses in corresponding panels of exhibit 10)

A) PLACE and MONTH EFFECTS and COMMON TERM

- East Coast, 56°F

  -- for Laredo, add 20°F
  -- for Washington, add nothing
  -- for Boston, subtract 4°F
  -- for Caribou, subtract 16°F
  -- for January, subtract 23°F
  -- for May, add 7°F
  -- for July, add 16°F

B) PLACE FITS and MONTH EFFECTS

- Laredo, 76°F
- Washington, 56°F
- Boston, 52°F
- Caribou, 40°F

  -- for January, subtract 23°F
  -- for May, add 7°F
  -- for July, add 16°F

C) MONTH FITS and PLACE EFFECTS

- January, 33°F
- May, 63°F
- July, 72°F

  -- for Laredo, add 20°F
  -- for Washington, add nothing
  -- for Boston, subtract 4°F
  -- for Caribou, subtract 16°F

D) PLACE and MONTH FITS, minus COMMON TERM

- January, 33°F
- May, 63°F
- July, 72°F

- Laredo, 76°F
- Washington, 50°F
- Boston, 52°F
- Caribou, 40°F

  -- to combine, add and then subtract 56°F from the sum.

exhibit 12 of chapter 16 -- East Coast temperatures

Effect and elementary analyses -- 3 each --
for our 3×4 example

A) EFFECT ANALYSIS -- in °F, based on exhibit 7.

	Laredo	Washington	Boston	Caribou	eff
January	6.3	1.0	.2	-7.5	-24.2
May	-.3	-.7	-1.3	2.1	6.0
July	-6.0	-.3	.9	5.4	18.2
eff	18.4	2.3	-4.0	-16.7	57.1

B) EFFECT ANALYSIS -- in °F, based on exhibit 8.

January	6.2	1.2	.5	-7.1	-24.2
May	-.5	-.6	-1.1	2.4	6.1
July	-6.0	0	-1.3	5.9	18.1
eff	18.7	2.3	-4.1	-16.9	56.9

C) EFFECT ANALYSIS -- in °F, based on exhibit 9.

January	4.6	-3.2	.1	-8.3	-23.0
May	-1.8	1.7	-1.2	1.5	7.0
July	-4.3	5.3	4.2	8.0	16.0
eff	20.0	0.0	-4.0	-16.0	56.0

D) ELEMENTARY ANALYSIS -- in °F, based on exhibit 7.

	Laredo	Washington	Boston	Caribou	eff	fit
January	6.3	1.0	.2	-7.5	-24.2	32.9
May	-.3	-.7	-1.3	2.1	6.0	63.1
July	-6.0	-.3	.9	5.4	18.2	75.3
eff	18.4	2.3	-4.0	-16.7	57.1	00.0
fit	75.5	59.4	53.1	40.4	00.0	-57.1

E) ELEMENTARY ANALYSIS -- in °F, based on exhibit 8.

January	6.2	1.2	.5	-7.1	-24.2	32.7
May	-.5	-.6	-1.1	2.4	6.1	63.0
July	6.0	0	-1.3	5.9	18.1	75.0
eff	18.7	2.3	-4.1	-16.9	56.9	00.0
fit	75.6	59.2	52.8	40.0	00.0	-56.9

F) ELEMENTARY ANALYSIS -- in °F, based on exhibit 9.

January	4.6	-3.2	.1	-8.3	-23.0	33.0
May	-1.8	1.7	-1.2	1.5	7.0	63.0
July	-4.3	5.3	4.2	8.0	16.0	72.0
eff	20.0	0.0	-4.0	-16.0	56.0	00.0
fit	76.0	56.0	52.0	40.0	00.0	-56.0

This is the case, now that a fit is something quite a little more complicated. We still cling to

data = fit + residual

and we still admit that there can well be different alternative fits to the same data.

To be sure, the last analysis in each set of 3 seems rather unsatisfactory, but it is AN analysis, too.

We will soon learn how to polish up analyses -- effect or elementary -- that we don't quite like. The ability to do this may make us more willing to live with having such diverse things called "an elementary analysis of this body of data".

* streamlining analysis by means *

The way we have done our analyses by means corresponds to starting with the most awkward assembly of our fit

fit by means = row mean + column mean - grand mean

and then changing fits to effects by writing this

fit = (row mean - C) + (column mean - C) + C

where

C = common term = grand mean.

We ought to be able to do the arithmetic a little easier by starting with one of the convenient (though unsymmetrical) two-term assemblies, such as

fit by means = (row mean) + (column effect) .

Exhibit 13 shows how this can be done by making our calculation in two steps. First we find, and subtract, row means. These means will be our row fits. Then we find column means of the result, and subtract them. They will be our row effects. What is left after the second subtraction are our residuals.

As carried out in exhibit 13, this process generates residuals, row effects, row fits, and column effects, leaving at most column fits to be calculated.

* parts of fits *

We need to stress that we can often understand a fit better in parts than as a whole. Exhibit 14 repeats the fit of exhibit 13 both ways. Look and ponder.

16E. Changing and combining fits; polishing

* a more general view *

The arithmetic we have just seen is helpfully thought of as two instances of a general procedure -- a half-step in which we start with a two-way fit and change it into another two-way fit we like better. Instead of starting with data alone, as we did in exhibit 13, we can quite properly say that we are starting with:

- the data, as residuals from a zero fit.
- a zero fit, divided into three parts, each of which is all zeros.

exhibit 13 of chapter 16 -- casual example

Two-step analysis by means

A) DATA, ROW SUMS, ROW MEANS

				sums	mean
58	7	-3	-104	-42	-10
2	-1	-9	1	-7	-2
-62	-2	7	28	-29	-7

B) DATA less ROW MEANS, also ROW FITS, both with COLUMN MEANS

	68	17	7	-94	-10
	4	1	-7	3	-2
	-55	5	14	35	-7
sum	17	23	14	-56	-19
mean	6	8	5	-19	-6

Examples: 68 = 58 - (-10); 4 = 2 - (-2)

C) SAME less their COLUMN MEANS, also COLUMN FITS

	62	9	2	-75	-4	(-10)
	-2	-7	-12	22	4	(- 2)
	-61	3	9	54	-1	(- 7)
(means)	6	8	5	-19	-6	

Examples: 62 = 68 - 6, -2 = 4 - 6;
-4 = -10 - (-6), 4 = -2 - (-6).

exhibit 14 of chapter 16 -- casual example

The fit of exhibit 13 looked at two ways

A) The FIT as an ASSEMBLED WHOLE

-4	-2	-5	-29
4	6	3	-27
-1	1	-2	-26

(What can you see at a quick glance?

B) The FIT in PARTS

				eff
.	.	.	.	-4
.	.	.	.	4
.	.	.	.	-1
6	8	5	-19	-6

(How much more can you see at a quick glance?)

Exhibit 15 starts just this way.

To be ready to take a step, we need not only to know where we are, we need to know how far to go -- and in which direction. In exhibit 13, the signposts were stripe means of residuals. In exhibit 15, they are still stripe means. For variety, however, we have chosen to work in the other order -- columns first, rows second. Rounding makes the result differ by ±1 in the second column, but otherwise all is the same.

Two things about the parts and the fit deserve notice:

- when we adjust a stripe of residuals one way, we adjust the corresponding part of the fit in the opposite direction by an equal amount. (This keeps our books balanced.)

- we calculate "next" values for one stripe of parts of fits as well as for all parallel stripes of residuals, we then subtract and add these part-of-fit "next" values just the same way as "next" values for stripes of residuals.

In the less important things, taking out MEANS is quite nice:

- after two half-steps, nothing remains to be done.

- either rows or columns can be taken out first, with the same result (also the same as taking out both together.

When neither resistance nor arithmetic labor matters -- a rather rare day -- it is nice to take out means. Usually we are led to do otherwise.

exhibit 15 of chapter 16 -- casual example

A general way to look at calculations like those of exhibit 13 -- modifying a two-way analysis (Here, stripe means of residuals come out next)

A) The INITIAL TWO-WAY ANALYSIS = the DATA -- and the INITIAL NEXT STEPS = COLUMN MEANS of RESIDUALS.

					part
	58	7	-3	-104	0
	2	-1	-9	1	0
	-62	-2	7	28	0
part	0	0	0	0	0
(next)	(0)	(1)	(-1)	(-25)	(0)

B) The FIRST half-step, "NEXT" SUBTRACTED FROM RESIDUALS and ADDED TO "PART" -- ROW MEANS of RESIDUALS and of "PART" row FORM new "NEXT"

					part	(next)
	58	6	-2	-79	0	(-4)
	2	-2	-8	26	0	(4)
	-62	-3	8	53	0	(-1)
part	0	1	-1	-25	0	(-6)

Examples: (2nd column) 6 = 7 - (-1), -2 = -1 - (-1), -3 = -2 - (1), 0 = 0 + (0), 1 = 0 + (1), -1 = 0 + (-1), etc.

C) ANOTHER HALF-STEP, again "NEXT" SUBTRACTED from RESIDUALS and ADDED to "PART" -- New "next" row all zero.

	62	10	2	-75	-4
	-2	-6	-12	22	4
	-61	-2	9	54	-1
part	6	7	5	-19	-6
(next)	(0)	(0)	(0)	(0)	

Examples: 62 = 58 - (-4), -2 = 2 - (4), -61 = -62 - (-1).
6 = 0 - (-6), 7 = 1 - (-6).
-4 = 0 + (-4), 4 = 0 + (4).
-6 = 0 + (-6).

D) REMARK: Panel C above differs from panel C of exhibit 13 by ±1 in 4 places (in column 2) because of differences in rounding. There would be no difference if arithmetic had been carried in exact fractions. (Preserve us from doing this!)

* working with medians *

This general procedure of adjustment by half-steps need not be confined to means. We can also use medians as guides to what changes to make next. Exhibit 16 does just this. It has to start somewhere, so panel C of exhibit 13, the close of a process of taking out means, was chosen as a start.

The calculations go along the same pattern. The only difference is that we use medians of stripes as "next" rather than means. After four half-steps, the process has not "closed" -- there is still one nonzero stripe median for residuals -- and one for parts of the fit.

As panel F shows, the called-for adjustments are getting smaller with decent speed. In larger tables, where there is more writing and subtraction, we usually -- and fortunately -- find adjustments getting smaller even more rapidly.

Since we often think of half-step adjustment by stripe medians as a way of fixing up analyses already made, it is natural to call the process

median polish.

We should not let this name persuade us that we need to begin with an almost satisfactory analysis. We will learn that one can often usefully begin median polishing untouched data -- data with a zero fit.

exhibit 16 of chapter 16 -- casual example

We can do the same thing with medians

A) The INITIAL ANALYSIS = panel C of exhibit 13 and COLUMN MEDIANS = DESIRED ADJUSTMENTS

					part
	62	9	2	-75	-4
	-2	-7	-12	21	4
	-61	-3	9	54	1
part	6	8	5	19	-6
(next)	(-2)	(-3)	(2)	(21)	(1)

B) SAME adjusted, with ROW MEDIANS = further DESIRED ADJUSTMENTS

					part	(next)
	64	12	0	-96	-5	(6)
	0	-4	-14	0	3	(-2)
	-59	0	7	33	0	(4)
part	4	5	7	40	-5	(6)

Examples: 64 = 62 - (-2), 0 = -2 - (-2), -59 = -61 - (-2)
12 = 9 - (-3), etc.
4 = 6 + (-2), 5 = 8 + (-3), etc.
-5 = -4 - (1), 3 = 4 - (1), (0) = 1 - (1).
-5 = -6 + (1).

C) READJUSTED -- with COLUMN MEDIANS OPEN for ADJUSTMENT

					part
	58	6	-6	-102	1
	2	-2	-12	2	1
	-63	-4	3	29	4
part	-2	-1	1	34	1
(next)	(2)	(-2)	(-6)	(2)	(1)

(cont'd.)

exhibit 16 (cont'd.)

D) READJUSTED AGAIN -- with ROW MEDIANS open

					part	(next)
	56	8	0	-104	0	(4)
	0	0	-6	0	0	(0)
	-65	-2	9	27	3	(4)
part	0	-3	-5	36	2	(-2)

E) READJUSTED YET AGAIN -- with COLUMN MEDIANS open

	52	4	-4	-108	4
	0	0	-6	0	0
	-69	-6	5	23	7
part	2	-1	-3	38	0
(next)	(0)	(0)	(-4)	(0)	(4)

F) SIZES of the SUCCESSIVE ADJUSTMENTS

|Sums of sizes of changes in parts|

Columns	Rows
28	
	12
12	
	3
4	

Examples: 28 = 2 + 3 + 2 + 21, from -2, -3, 2, 21.
12 = 6 + 2 + 4, from 6, -2, 4.

G) PROBLEM

1) Carry out this process, beginning with the analysis implied by exhibit 9.

* how much polish? *

If we have to fix a standard for row to median polish, we probably do well to choose four half-steps as the standard. We do well, indeed, to think of 4 as a typical target, with the analyst free to use fewer half-steps whenever they seem enough, two do quite frequently -- and allowed to use more in the rare cases where this seems worth while. (In a later chapter we will learn of a convenient modification for larger tables.)

* other kinds of polish *

We know about other central values besides medians and means. We can use trimeans -- or midmeans -- or what you will. Nothing essential changes. The mean -- in its coming to completion after a fixed number (in its case 2) of half-steps -- appears to be unique. All the others behave like the median, running on for varying numbers of steps.

We are free to use these others when we will. In a 3×4 example, there is going to be little to gain from many of them, since they will reduce, in batches of 3 or 4, either to the median or to the mean. (Skipping central values are exceptions, of course.

* combining analyses *

Sometimes the data with which we start a polishing calculation are themselves residuals in a previous analysis, as exhibit 17 illustrates.

In this situation we combine the analyses as follows:

- copy residuals from the inner analysis.
- find parts of fit as sums of the corresponding parts of outer and inner analyses.

16F. Checking two-way analyses

* the need for checking *

Two-way analyses, even if done as neatly as we know how, involve 4 additions or subtractions per data value (two half-steps using means or four half steps using medians). By hand, arithmetic is rarely without error unless checked. (And not always then.) We need to have a good check of polishing, whether done by means, by medians, or how.

What ought we check? There are real choices, including:

- are we happy with where we stopped?
- is what we have done possible -- does it in fact analyze the data with which we started?
- did we do every step along the way correctly?

These three deserve quite different sorts of attention:

exhibit 17 of chapter 16 -- East-coast temperatures

Nested analyses
(back to °F)

A) The ORIGINAL DATA, with a NULL FIT -- see exhibit 7

	Laredo	Wash'n	Boston	Caribou	part
January	57.6	36.2	29.1	8.7	.0
May	81.2	64.7	57.8	48.5	.0
July	87.7	77.3	72.2	64.0	.0
part	.0	.0	.0	.0	.0

B) The OUTER ANALYSIS, see exhibit 9 for basis

					part
January	4.6	+3.2	.1	-8.3	-23.0
May	-1.8	1.7	-1.2	1.5	7.0
July	-4.3	5.3	4.2	8.0	16.0
part	20.0	.0	-4.0	-16.0	56.0

C) The NEWLY ASSUMED DATA

					part*
January	4.6	3.2	.1	-8.3	.0
May	-1.8	1.7	-1.2	1.5	.0
July	-4.3	5.3	4.2	8.0	.0
part*	.0	.0	.0	.0	.0

D) The INNER ANALYSIS -- MEAN POLISHED ANALYSIS for NEW DATA

					part*
January	6.3	1.0	.3	-7.5	-1.2
May	-.3	-.7	-1.2	2.1	-1.0
July	-6.0	-.3	1.0	5.4	2.2
part*	-1.6	2.3	-.1	-.7	1.1

E) RESULT of COMBINING panels B and D

					part
January	6.3	1.0	.3	-7.5	-24.2
May	-.3	-.7	-1.2	2.1	6.0
July	-6.0	-.3	1.0	5.4	18.2
part	18.4	2.3	-4.1	-16.7	57.1

Notes: Residuals copied from panel D.
Part values are sums of "part" from panel B and "part*" from panel D.

Values differ from those found in exhibit 7 by ±0.1 in the 3rd column; caused by rounding.

- the first has to be up to the higher judgment -- sometimes the highest judgment -- of the analyst. No mechanical process can deal adequately with the question "Does this seem to be a satisfactory analysis?"

- the second can be dealt with mechanically -- and rather easily -- as we shall soon see. It has to be given real attention, however, for there is little profit in analyses that are not of the right data.

- the third is something to avoid in an exploratory situation. If the tests of satisfactoriness and possibility are met, just how we got there is rather secondary. (In confirmatory data analysis, where what would have happened had the data been different is always relevant, we are likely to need to worry about exactly how we reached a result -- but not in exploratory data analysis.)

* making the check *

So let us plan to check only the possibility of a supposed modified analysis. If two two-way analyses analyze the same data, their difference -- another two-way analysis -- will analyze zero. Our problem reduces to:

- finding a difference analysis.

- checking to see if the difference analysis really "analyzes zero".

Panel A of exhibit 18 finds the difference analysis between the analyses of panels A and E of exhibit 15. The check-sum stripes are sums of the two nearest values -- with one exception each (in bold characters) one of the two values is a part of the fit, the other a residual. These check sums turn out to be constant in our example, -12 for the bottom check-sum stripe, -25 for the right-hand check-sum stripe.

This constancy is the first step in checking that our analysis does analyze zero. What it checks is this: Are the last row and column of whatever is analyzed all the same? In the last row, for instance, what is analyzed is:

-8 + (-4) + (row part) + (common)

-3 + (-9) + (row part) + (common)

-4 + (-8) + (row part) + (common)

-31 + (19) + (row part) + (common)

By checking that the changing terms add to -12 in each case, we have checked that all these 4 values are exactly the same.

To be sure that they are zero, we need to know that

(row part) + (common)

is equal and opposite to -12. The last (bold) entry in the check-sum row checks this. Interchanging rows and columns affects nothing. As a result our check-sum stripes have checked that the last row AND last column of what is analyzed are made up of all zeros.

exhibit 18 of chapter 16 -- casual example

Checking the possibility of panel E of exhibit 16 as an analysis of the same data as panel A of that exhibit

A) DIFFERENCE ANALYSIS with CHECK-SUM STRIPES -- panel panel E MINUS panel A (both exhibit 16)

					part	check-sums
	-10	-5	-6	-33	8	-25
	2	7	6	-21	-4	-25
	-8	-3	-4	-31	6	-25
part	-4	-9	-8	19	6	25
check-sums	-12	-12	-12	-12	12	okay

Examples (parenthetic entries from panel A, others from panel E of exhibit 16):

-10 = 52 - (62), 2 = 0 - (-2), -8 = -69 - (-61).
-4 = 2 - (6), -9 = -1 - (8).
8 = 4 - (-4), -4 = 0 - (4).
6 = 0 - (-6).

(check-sums)

-12 = -4 - 8, -12 = -9 - 3, ..., 12 = 6 + 6.
-25 = -33 + 8, -25 = -21 - 4, ..., 25 = 19 + 6.

B) LOOKING at DIFFERENCES of ADJACENT RESIDUALS -- the STRIPE CONSTANCY CHECK.

-10	-5	-6	-33	
✓	✓	✓	✓	(all +12)
2	7	6	-21	
✓✓	✓✓	✓✓	✓✓	(all -10)
-8	-3	-4	-31	

Check of possibility finished.

Panel B shows how we spread the check to the rest of the table. We look at differences of adjacent entries to see if these are properly constant. The easy way is usually to difference down columns and check constancy across rows, as we have done in panel B. It suffices to make enough checks of this sort to involve every residual in one difference, but it often pays to go further.

The check of panel B -- the stripe-constancy check -- is usually best dealt with before worrying about the check-sum stripes. We may find, by checking among the residuals, that a residual or two in the last row or column needs to be corrected. Fixing it will then tend to fix the corresponding entry in a check-sum stripe.

The ability to make the stripe-constancy check using differences either within columns or within rows is important. Using a particular one often makes it easy to pin-point an error -- and then correct it.

We will have further examples and discussion of checking when we come to deal with larger tables.

* correct, and go on *

A great virtue of this checking procedure is just its ability to find what fix is needed to make the analysis we were checking a possible one -- to make it an analysis of the data we started with. We will ordinarily use this ability, make the fix, and then go back to the first question "is the analysis satisfactory."

There is, in exploratory data analysis, almost never a need to go back and find where the error happened. If the corrected analysis seems satisfactory, fine -- we can stop there. If not, we can go on, modifying the analysis in the best way we can see.

Checking should promote forward progress, not just backward rectification.

16G. Making two-way plots easier and more effective

* easier construction *

The pictures with which we began this chapter are nice to look at, but -- if we do not know the trick -- they can be messy to draw. Ordinary graph paper does not seem planned to draw rectangles balanced on their corner.-- at least until we turn the graph paper through 45°, when we have the convenient relation

fit = row effect + column fit.

If we take row effect and column fit as the two axes on our graph paper, then

fit = constant

is a 45° line, since it is of the form

$$x+y = \text{constant}.$$

exhibit 19 of chapter 16 -- casual example

Median polishing stripe effects,
changing the common term to match

A) The STARTING ANALYSIS -- from exhibit 13, panel A

					eff
	58	7	-3	-104	6
	2	-1	-9	1	-3
	-62	-2	7	28	3
eff	-1	-3	-1	23	0

Median of row effects: 3
Median of col effects: -1
Sum of these: 2

B) The result of MEDIAN POLISHING the EFFECTS

58	7	-3	-104	3
2	-1	-9	1	6
-62	-2	7	28	0
0	-2	0	24	2

Note how clearly the right-most column effect sticks out.

C) The CONVENTIONAL form of the SAME ANALYSIS -- as in exhibit 13B, except still in 0.1°F

					eff
	58	7	-3	-104	4
	2	-1	-9	1	-5
	-62	-2	7	28	1
eff	-5	-7	-5	19	6

Note how much less obtrusive -- size 19 against a nearest competitor of size 7 -- the rightmost column effect seems.

D) CALCULATIONS for MEDIAN POLISHING the EFFECTS in the CONVENTIONAL form

Median of row effects: 1
Median of col effects: -5
Sum of these: -4
Common term: 6
Sum of last two: 2

Thus, to median polish effects, we should subtract 1 from each row effect, add 5 to each column effect, and compensate by subtracting 4 from the common term. This leads us right back to panel B, as it should.

All we need do, then is to draw our rectangular diagram of the fit on these two axes, mark in lightly a few lines $x+y$ = constant and turn the graph paper to make these light lines horizontal before tracing the result on the tracing paper we were going to use anyway. As we trace, we may replace the light horizontal lines by a vertical scale. Exhibit 20 shows the fit for our 3×4 example as drawn on the graph paper. Exhibit 21 shows the result as traced. Given the effects and fits, making this kind of plot is easy.

* showing outside residuals *

We also want to show carefully selected information about residuals. We begin with those residuals that are outside -- when looked at in the batch of all residuals. Where do they appear? What is their direction -- down or up? How big are they? Also -- how big might a residual be that is not outside? What is a useful summary of the other residuals?

Construction plot, as drawn on graph paper

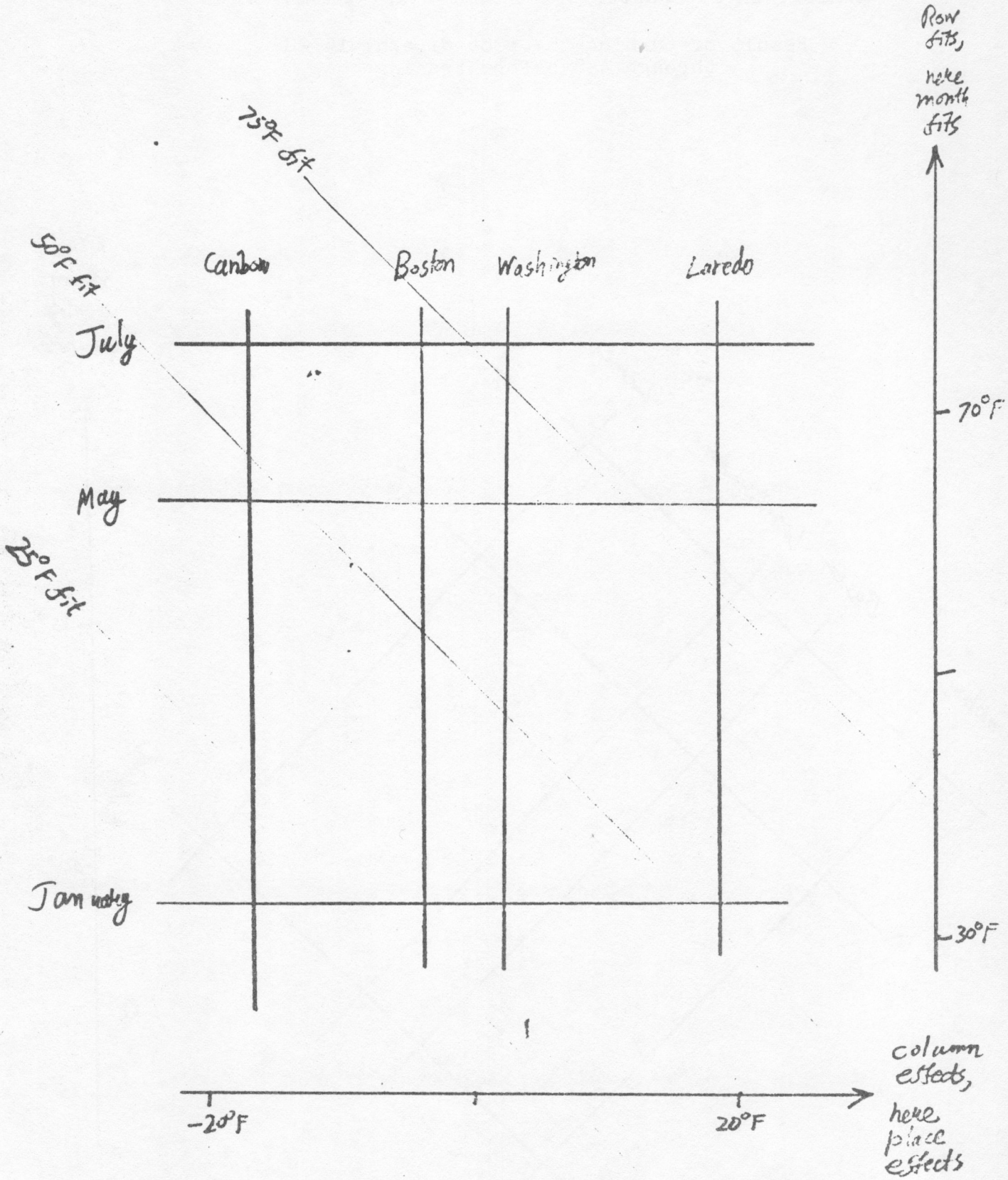
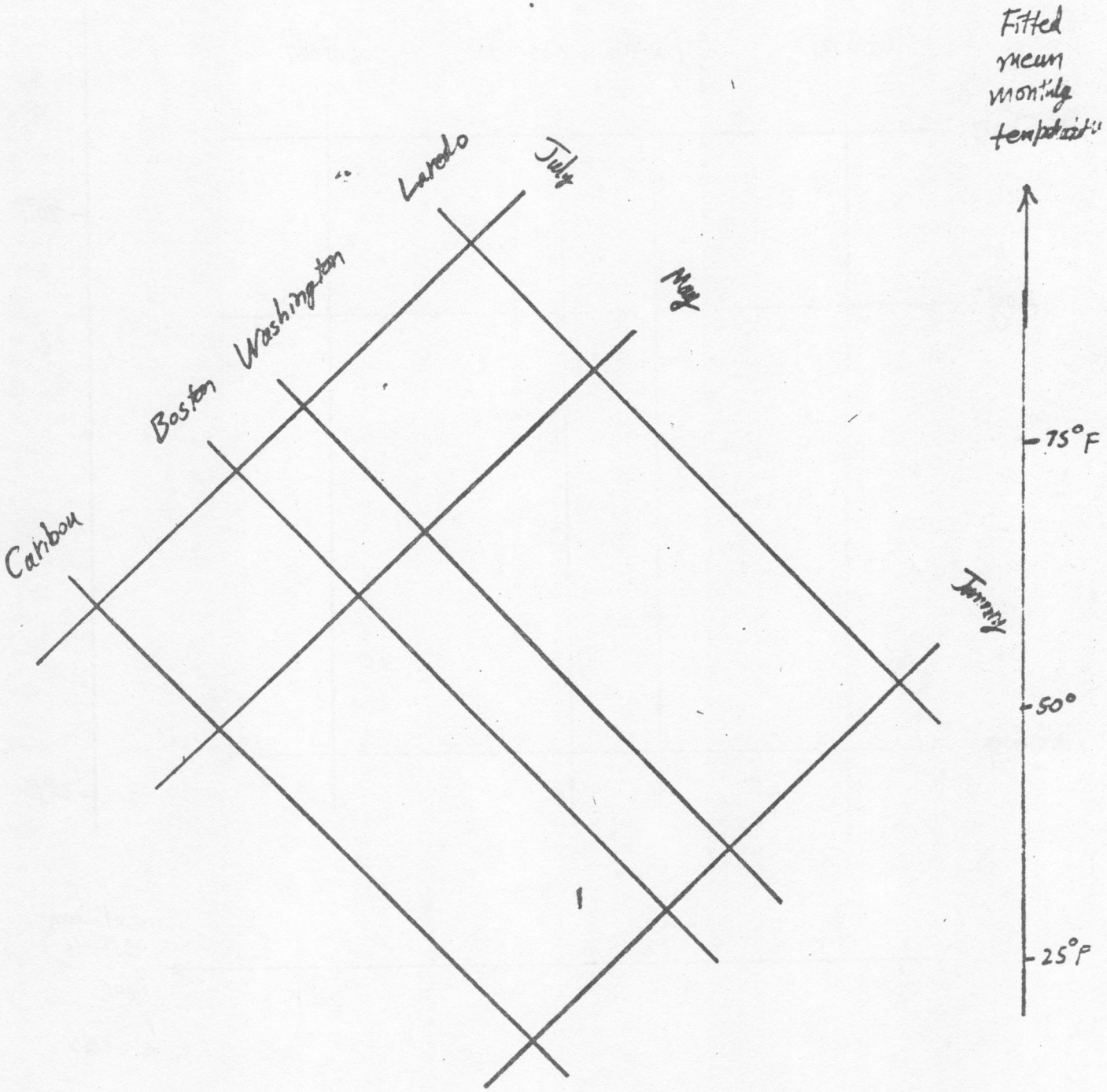


exhibit 21 of chapter 16 -- East-coast temperatures

Result of turning the plot of exhibit 20
through 45° before tracing



There are usually few enough residuals outside -- beyond the sides -- so that we can show them on the two-way plot. The natural way to do this, as we saw in exhibit 5, is to draw arrows from "fit" to "fit plus residual". These arrows run at 45° on the construction plot, and are vertical on the two-way plot. An easy way to draw them is to put an end-line on the construction plot at the right height for their end, and then draw the vertical arrow ending on this end-line while tracing the two-way plot.

If the residual is +5.4, as it is for Caribou in July (see exhibit 7), we start at the Caribou-July intersection on the construction plot and move up (vertically) the Caribou line 5.4°F and plot a point. Then we start at the same intersection and move up (to the right) the July line 5.4°F and plot another point. The line joining these two new points corresponds to a temperature 5.4°F higher than the intersection and thus is the desired end-line. This construction is then repeated for all the other outside residuals.

* summarizing other residuals *

What about the adjacent and inside residuals? We are summarizing, so it is natural to use one or more schematic plots. A convenient way to do this is to use a strip at the bottom of the picture, both to show how large the other residuals can be and to remind us of the relevant numbers. Exhibit 23 shows how this can be done for our 3×4 plot of East Coast temperatures.

exhibit 22 of chapter 16 -- East-coast temperatures

Coding the residuals for our 3×4

A) RESIDUALS from panel F, exhibit 7, DISPLAYED

|High|	|Low|					
6.4	-6.0	2	H	HH	4 + 2 + 6 = 12, hinges at 3h	
5.4	-7.5	3	2	1		
		4	1	0	Hinges:	+1.6 and -0.5
		(2)	0	29	Midspread:	2.1
		6	-0	3733	Sides:	+3.7 and -2.6
			-1		Outside	2 and 2
		2	L	LL	(count/20)	(h)

B) OUTSIDE RESIDUALS ARRAYED

	Laredo	Washington	Boston	Caribou
January	+6.3	.	.	-7.5
May	.	.	.	.
July	-6.0	.	.	+5.4

C) CODED RESIDUALS ARRAYED -- see text

January	X	.	.	0
May	.	o	.	x
July	0	.	.	X

* looking harder at the pattern *

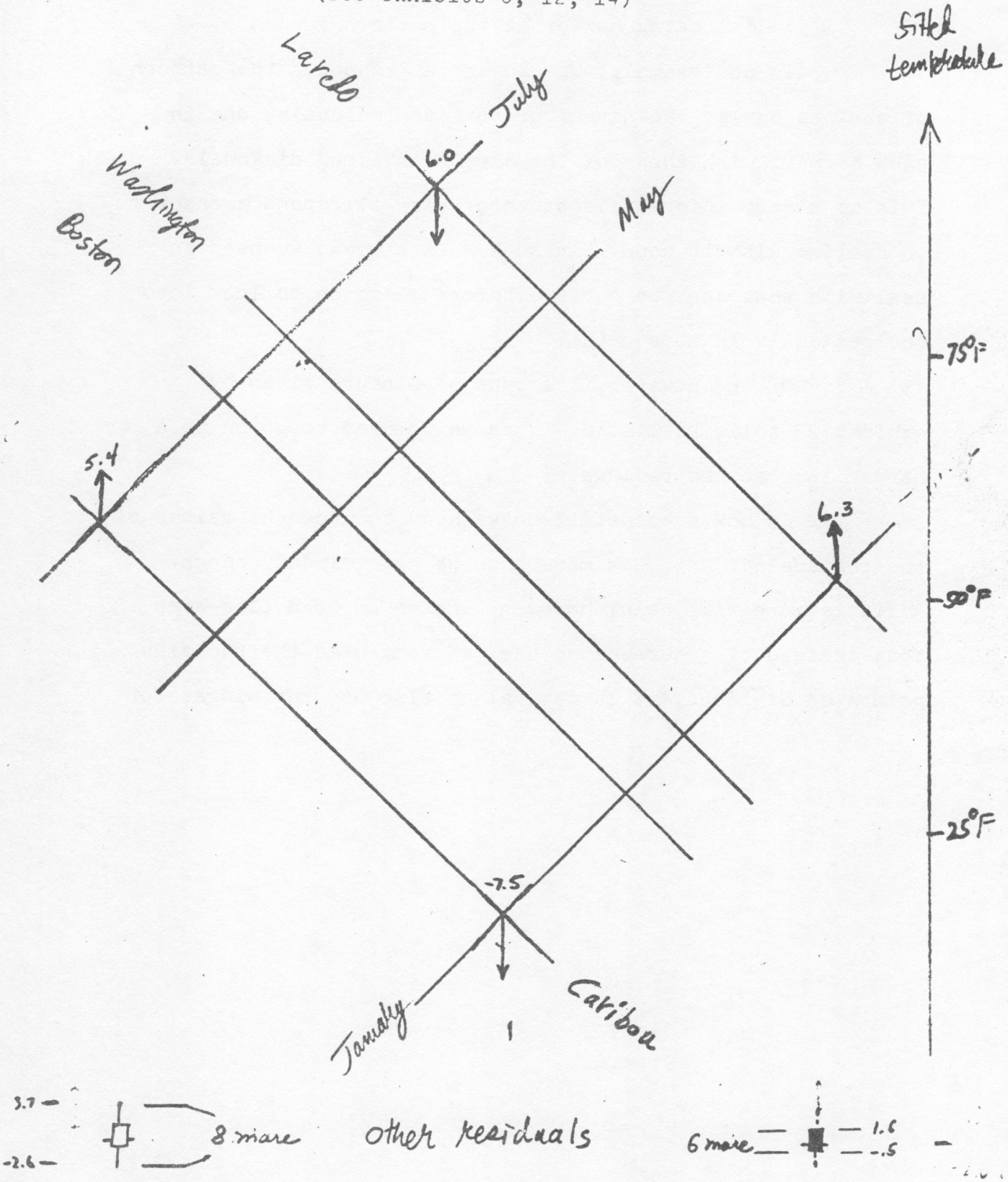
In our example, we can see a lot about the pattern of our residuals. We have four outside residuals, one in each corner, with those of the same sign along diagonals. This is a very informative pattern; we will spend a chapter on dealing with it soon. In such a situation, we need to deal with what strikes our eye before we go on to look into the residuals in more detail.

Often, however, the type of picture given by exhibit 23 tells us little. Then we do need to go on to a harder look at the residuals.

In how much detail do we need to know the values of the residuals? Not much more than we have already shown. A little more will help, however. So we go to a five-step code instead of a three-step one. Having used the two side points as dividers, it is natural to also use the hinges for

exhibit 23 of chapter 16 -- East-coast temperatures

Graphic portrayal of our analysis of the 3×4
(see exhibits 6, 12, 14)



this purpose. We will routinely use the following code:

large X

(upper side point)

small x

(upper hinge)

dot

(lower hinge)

small o

(lower side point)

large O

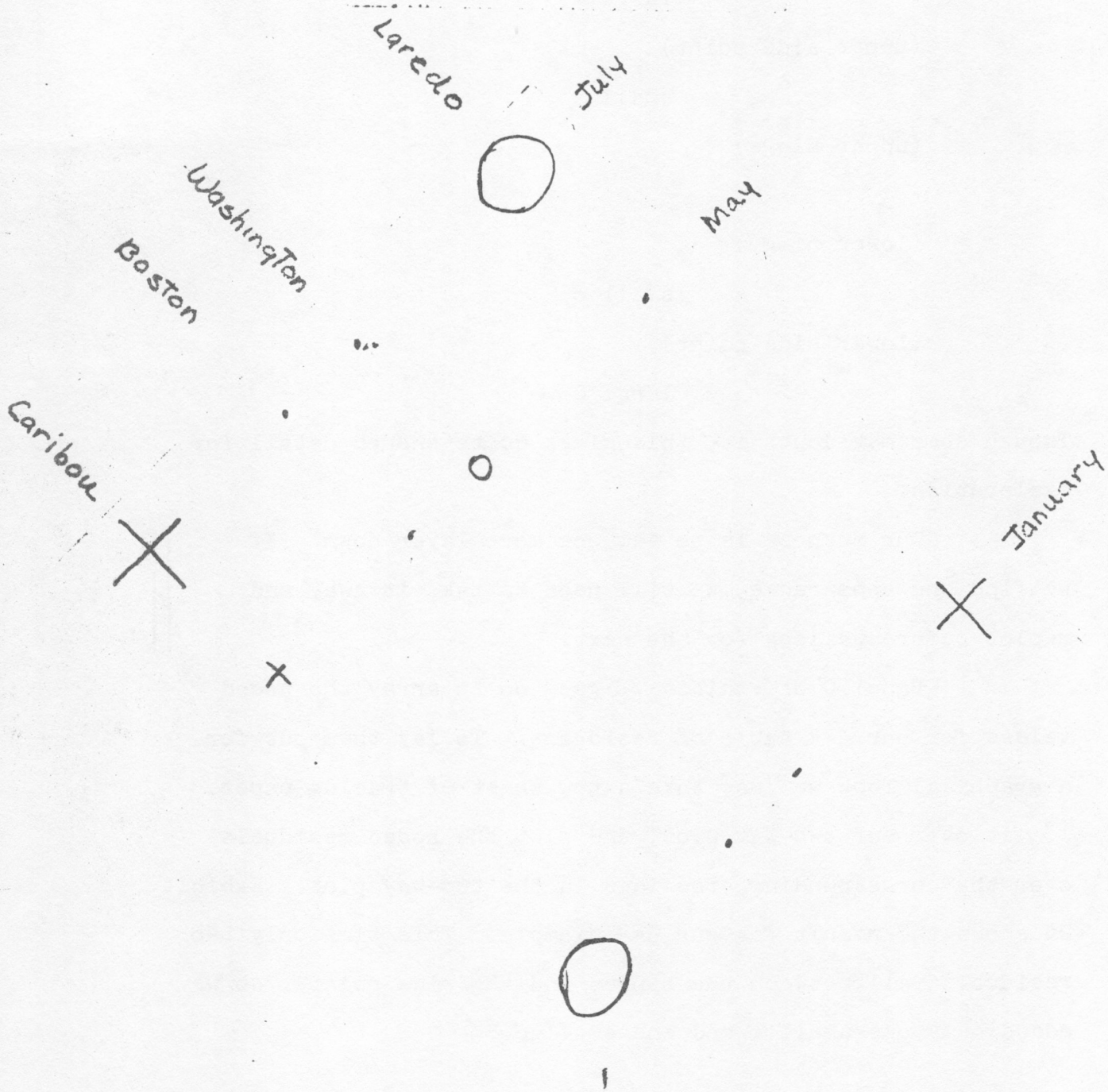
Though some may doubt it, this gives quite enough detail for exploration.

Our purpose is to see one more layer down. If we find one appearance, we will need to take it away and replot before we look for the next.

Panel C of exhibit 22 goes on to array the coded values for our 3×4 table of residuals. To lay them out for a graphical look we just take a new sheet of tracing paper, lay it over our two-way plot, and plot the coded residuals over the corresponding crossings in the two-way plot. Exhibit 24 shows the result for our 3×4 example. This time only two residuals fall between the hinges and the side points, so we added only one small x and one small o.

exhibit 24 of chapter 16 -- East-coast temperatures

Looking harder at the pattern of residuals



Notice that nothing more, except minimum labels lightly fenced off, is shown on this picture. If we want full freedom to make good use of this plot, it is important to put unneeded detail away. Examples with larger numbers of residuals will show this to us in due course.

Even in this case, however, we can see a tendency for May to behave a little like July -- o's to the NE and x's to the SW -- though not at all so strongly.

Sometimes we can usefully show the two plots side-by-side. Exhibit 25 shows how this can look.

exhibit 25 of chapter 16 -- East-coast temperatures

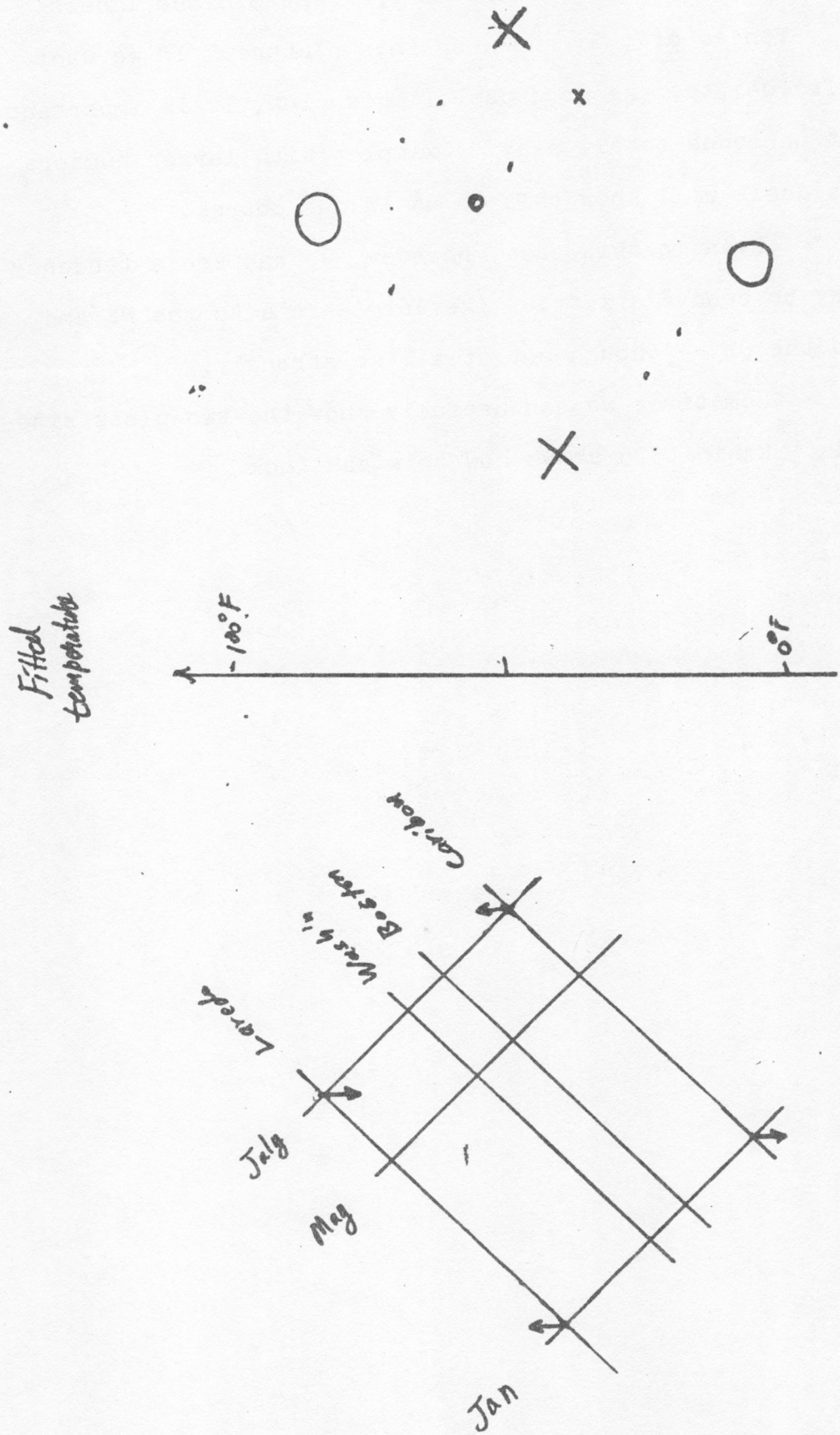


exhibit 25 of chapter 16

The two plots shown side by side

****

16H. <u>How far have we come</u>?

If this chapter has done its work, we know -- and are beginning to understand:

- what two-way tables are.
- what the results of their simplest analyses look like.
- something about how to think of these analyses.
- some ways to calculate, modify, and combine them.
- how to check that two-way analyses are not inconsistent.
- some ways to plot both fits and residuals so that we can see what appears to be going on.

Two-way tables tell us a response for combinations of a version of one factor and a version of another.

Their simplest analysis follows our old pattern

data = fit + residual .

The fit can be written as a sum of simple parts, for example:

fit = column effect + row effect + common

fit = column effect + row fit

fit = column fit + row effect

The fit is simpler to understand in terms of parts than as a whole. We can, if we wish, show all three of these sets of parts in one elementary analysis. We find it easiest to

calculate, modify, and combine effect analyses that give us only the three-part breakdown above.

Once we are used to two-way tables, any two-way table of data almost seems to have a ghostly zero fit around it. We rarely want to stop there.

We can calculate a two-way fit by means. The arithmetic labor is reasonable for small tables, up to 4×4. The resistance of the result is poor.

We can get a resistant result by median polishing -- starting with either a first analysis or the raw data. The individual stage in median polishing is a half-step. Four half-steps are very often enough. Two or three are frequently enough.

Checking hand arithmetic is increasingly necessary as the structure of the data becomes more complicated. Checking that a two-way table analyzes what is supposed to is easy. A difference analysis, that should analyze zero, can be checked in terms of a stripe constancy check on differences of residuals, and a check-sum check combining residuals and effects. We ought to make such checks routinely. In exploratory data analysis, we should fix any impossibilities in the result to be checked, and go right on without looking to see where the error was made.

We can plot any fit of the form considered in the chapter quite simply. We find a rectangular array of 45° lines, one for each version of each factor. The value of the fit corresponding to any combination of versions is shown by the height of the point at which the corresponding lines intersect. The left-to-right position of this intersection is ordinarily to be forgotten about. (Ordinary graph paper, used normally -- and then turned through 45° -- lets us make such plots easily.)

It pays to show some information about certain residuals on such a plot. If we treat all residuals as a batch, some residuals (or none) will be outside. These we mark on the two-way plot. We also show the general sizes of the unidentical residuals at the foot of the picture.

Sometimes we need to look harder at the residuals. A five-group coding, X,x,·,o,O suffices. We plot these coded values on a separate sheet of tracing paper, placing each one according to the corresponding intersection on the two-way plot of the fit.

We have begun to grasp the basic ideas about two-way tables -- we have much still to learn.

****

16K. Median polishing without choices (background)

We have combined the introduction of median polishing with a rather free and easy attitude to two-way analysis. While this is the attitude most appropriate to exploratory data analysis, there are purposes for which it is less appropriate. It would be wrong to carry over toward these other purposes the idea that median polishing has to be free and easy. There can be rigid, standardized analyses for two-way tables that depend on medians for resistance, as we shall now see.

In the easy and effective median analyses we use routinely, one choice has to be whether to adjust for row medians before adjusting for column medians or vice versa. The way out is to adjust for both together -- taking a full step instead of two half-steps.

* full-step polishing *

Exhibit 26 shows how this proceeds. The first full step shows nothing unusual. The medians are combined so that we may adjust for both at once.

Since combining the medians allows for the general size of the data twice, the values after the first step are, as a batch, somewhere nearly minus the original values. When we take the second full step, we combine halves of medians to avoid coming back where we started. (This tends to have

exhibit 26 of chapter 16 -- coronal height

Taking a full step in median polishing

A) The DATA -- HEIGHT of the SOLAR CORONA

Solar activity	Angular distance from Solar Equator				(median)
	0°	30°	60°	90°	
Minimum (1953-54)	1.99	1.41	.57	.62	(1.02)
Rising (1955-56)	1.34	1.62	.65	.63	(1.00)
Maximum (1957	1.40	1.76	.73	.56	(1.06)
(median)	(1.40)	(1.62)	(.65)	(.62)	

B) COMBINATION of MEDIANS

	1.40	1.62	.65	.62
1.02	2.42	2.64	1.67	1.64
1.00	2.40	2.62	1.65	1.62
1.06	2.46	2.68	1.71	1.68

C) FIRST FULL STEP -- panel A minus panel B

					(median)
Min	-.43	-1.23	-1.10	-1.02	-1.06
Rise	-1.06	-1.00	-1.00	-.99	-1.00
Max	-1.06	-1.08	-.98	-1.12	-1.07
(median)	-1.06	-1.08	-1.00	-1.02	

D) COMBINATION of HALVES of MEDIANS

	-.53	-.54	-.50	-.51
-.53	-1.06	-1.07	-1.03	-1.04
-.50	-1.03	-1.04	-1.00	-1.01
-.54	-1.07	-1.08	-1.04	-1.05

exhibit 26 (cont'd.)

E) SECOND STEP -- panel C minus panel D

Min	.63	-.16	-.07	.02
Rise	.03	.04	.00	.02
Max	.01	.00	.06	-.07

F) CALCULATION of EFFECTS

Step	0°	30°	60°	90°	(Medn)	**Min**	**Rise**	**Max**	(Medn)
1st	1.40	1.62	.65	.62		1.02	1.00	1.06	
2nd	-.53	-.54	-.50	-.51		-.53	-.50	-.54	
Sum	.87	1.08	.15	.11	(.51)	.49	.50	.52	(.50)
eff	.36	.57	-.36	-.40		-.01	.00	.02	

Examples: .87 = 1.40 + (-.53), .36 = .87 - .51.

G) The TWO-WAY ANALYSIS

Min	.63	-.16	-.07	.02	-.01
Rise	.03	.04	.00	.02	.00
Max	.01	.00	.06	-.07	.02
eff	.36	.57	-.36	-.40	1.01

Example: 1.01 = .51 + .50 .

H) SOURCE:

M. Waldemeier 1958. Die form der Korona nach Beobacttungen in der Linie 6374Å. Zeitschrift für Astrophysik 45: 155-170. From table 3 on pages 164-165.

16-63

other advantages also.) In this example, the results is a reasonable, but not strikingly good analysis.

* a possible standard *

One standard two-way analysis that might be chosen runs as follows:

- make an analysis by means.
- polish that analysis by 2 full steps, the first based on medians, the second based on halves of medians.

Exhibit 27 carries this through for the solar corona data, doing the arithmetic in half-step form but in such a way as to reach the full-step results. Again the result is a reasonable but not strikingly good analysis.

exhibit 27 of chapter 16 -- coronal height

Mean analysis followed by twin-median analysis

A) The DATA -- see exhibit 26

	0°	30°	60°	90°	(mean)
Min	1.99	1.41	.57	.62	(1.15)
Rise	1.34	1.62	.65	.63	(1.06)
Max	1.40	1.76	.73	.56	(1.11)

B) ADJUSTED

					part
Min	.84	.26	-.58	-.53	1.15
Rise	.28	.54	-.41	-.43	1.06
Max	.29	.65	-.38	-.55	1.11
(mean)	(.47)	(.48)	(-.46)	(-.50)	(1.11)

C) AGAIN ADJUSTED -- mean analysis

					eff	(median)
Min	.37	-.22	-.12	-.03	.04	(-.08)
Rise	-.19	.06	.05	.07	-.05	(.06)
Max	-.18	.17	.08	-.05	.00	(.02)
eff	.47	.48	-.46	-.50	1.11	(.00)
(median)	(-.18)	(.06)	(.05)	(-.03)	(.00)	

D) ADJUSTED for COLUMN MEDIANS from panel C

Min	.55	-.28	-.17	.00	.04
Rise	-.01	.00	.00	.10	-.05
Max	.00	.11	.03	-.02	.00
eff	.29	.54	-.41	-.53	1.11

E) ADJUSTED for ROW MEDIANS from panel C

					eff	(median)
Min	.63	-.20	-.09	.08	-.04	(.00)
Rise	-.07	-.06	-.06	.04	.01	(-.06)
Max	-.02	.09	.01	-.04	-.02	(.00)
eff	.29	.54	-.41	-.53	1.11	(-.06)
(median)	(-.02)	(-.06)	(-.06)	(.04)	-.02)	

exhibit 27 (cont'd.)

F) ADJUSTED for HALVES of COLUMN MEDIANS from panel E -- full adjustment for effects.

					eff
Min	.64	-.17	-.06	.06	-.02
Rise	-.06	-.03	-.03	.02	.03
Max	-.01	.12	.04	-.06	.00
eff	.29	.54	-.41	-.53	1.09

G) ADJUSTED for HALVES of ROW MEDIANS from Panel E -- full adjustment for effects

					eff
Min	.64	-.17	-.06	.06	-.02
Rise	-.03	.00	.00	.05	.06
Max	-.01	.12	.04	-.06	-.02
eff	.35	.60	-.35	-.59	1.03

H) PROBLEMS

1) Starting from panel G repeat the twin-median polish, as from panel C to panel G.

2) Starting from panel C, do the first full step as in exhibit 26. How does the result compare with panel E?

3) The same for panel E to panel G.

4) Alter the upper-left data entry from 1.99 to 13.99 and repeat all calculations. What can you say about the relative resistance of the analyses of panels C and G?

****

16P. Problems for chapter 16.

1. Do problem 1 of exhibit 7.

2. Do problem 2 of exhibit 7.

3. Do problem 3 of exhibit 7.

4. Do the problem of exhibit 16.

5. Do problem 1 of exhibit 27.

6. Do problem 2 of exhibit 27.

7. Do problem 3 of exhibit 27.

8. Do problem 4 of exhibit 27.

9. J. M. Hartwell, Jr. (1947. Interest rates by loan size and geographical region. Journal of the American Statistical Association 42: 242-245) reported average rates of interest (% per year) for loans of sizes ($100, $700, $1000, $3,000, $10,000) in three locations as follows: New York City (5.0,4.0,3.5, 3.3, 2.8). South and West (5.8, 5.4, 4.8, 4.1, 3.6). North and East (5.9, 5.3, 5.2, 4.1, 3.6). Make a two-way plot with appropriate plots. Comment on the results.

10. W. J. Youden (1946. Technique for testing the accuracy of analytical data. Analytical Chemistry 19: 946-950. His table VI on page 949) gives measurements of the voltages between an antimony and a hydrogen electrode at various pH's. The numbers following are for set (a,b) and are excesses over 255.00 millivolts given in units of 0.01 millivolts. At pH 2.20 (38, 34). At pH 3.00 (12, 12). At pH 4.20 (9, 10). At pH 5.00 (1, 0). At pH 6.00 (61, 41). At pH 6.80 (91, 74). At pH 8.00 (171, 172). Make a reasonable two-way analysis, as resistant as you can. Make all helpful plots. Comment on the result.

11. M. Smith (1948. Occupational differentials in physical status. American Sociological Review 13: 72-82. His table 5 on page 80.) gave the following rates of physical defects per 1000 registrants for the World War II draft for (eyes, teeth, hernia, feet) by occupation. Professional and semi-professional (160,102,56,95). Proprietors, managers and officials (145,127,56,100). Clerical and sales (158,128,55,91). Domestic service workers (92,107,48,120). Farm laborers and foremen (93,79,59,97). Students (157,75,33,75). Make a two-way analysis, with all appropriate plots.

12. Holmes and Robbins (1946. Rubber determination in young guayule. Studies on the Spence-Caldwell method. Analytical Chemistry 19: 313-317. Their table I on page 314) give % of rubber for 9 samples after holding (0, 3, 24, 48) hours at 100°C, namely: Sample A (3.80, 3.74, 2.90, 1.51). Sample B (3.20, 3.15, 2.25, 1.09). Sample C (c.70, 3.80, 3.43, 1.27). Sample D (4.04, 3.65, 3.19, 1.14). Sample E (2.99, 3.01, 2.72, 0.95). Sample F (3.65, 3.36, 2.86, 2.88). Sample G (3.89, 3.51, 2.83, 2.46). Sample H (4.38, 4.15, 3.78, 3.52). Sample I (12.73, 12.70, 10.79, 8.68). Make a reasonable resistant two-way analysis, including plots.

13. A. Editor (1947. The 1947 census of research projects. A. Introductor Statement. American Sociological Review 12: 448-449) reported the following numbers of projects in sociology in (1945, 1946, 1947): Social psychology (62, 87, 92). Population (31, 35, 62). Social problems (34, 40, 55). Marriage and family (37, 40, 54). Social research (23, 38, 38). Make a two-way analysis with appropriate plots. Comment on the results.

14. Harban and McGlamery (1963. Limitations on measuring flow rates of polyethylene and ethylene copolymers by extrusion plastometer. Materials Research Standards 3: 906-909. Their table 2 on page 907) give the flow rate of samples (A,B,C,D,E) in grams per 10 minutes according to different methods of measurement. Method 1 (0.26, 1.70, 3.60, 4.88, 6.80). Method 2 (17.8, 98, 189, 112, 283). Method 3 (1.54, 10.80, 25.12, 27.48, 37.22). Make a reasonable two-way analysis with plots. Discuss the results.

15. Moyneaux, Gilliam and Florant (1947. Differences in Virginia death rates by color, sex, age, and rural or urban residence. American Sociological Review 12: 525-535) give the following age-spectric death rates (per 1000) for (rural males, rural females, urban males, urban females): Ages 50-54 (11.7, 8.7, 15.4, 8.4). Ages 55-59 (18.1, 11.7, 24.3, 13.6). Ages 60-64 (26.9, 20.3, 37.0, 19.3). Ages 65-69 (41,0, 30.9, 54.6, 35.1). Ages 70-74 (66.0, 54.3, 71.1, 50.0). Round these values to the nearest unit and make a two-way analysis, with appropriate plots. Comment on the results.

Apply the fit of the last problem to the full data in units and tenths, and median polish the results. Comment on the changes.

16. Goetz, Loomis and Diehl (1950. Total hardness in water. Stability of standard disodium dihydrogen ethylenediamine tetraacetate solutions. Analytical Chemistry 22: 798-799. Their table II on page 799) give the strength of standard versenate solutions after different times of storage at different acidities or alkalinities (different pH's). The numbers following are the excess -- in units of 0.001 -- of the strength of solution in milligrams equivalent of calcium carbonate ($CaCO_3$) per milliliter over 1.000 after (0, 5, 30, 60, 120) days. At pH5 (49, 48, 49, 47, 42). At pH6 (44, 44, 42, 41, 35). At pH7 (62, 63, 60, 58, 51). At pH8 (50, 50, 47, 42, 34). At pH9 (62, 58, 43, 30, 19). At pH10 (47, 41, 19, 2, -30). Make a reasonable, resistant two-way analysis, with all helpful plots. Discuss the result.

17. R. L. Anderson (1947. Use of variance components in the analysis of hog prices in two markets. Journal of the American Statistical Assocation 42: 612-634) gave the following comparisons of the price of hogs (180-200 pound class) in Cincinnati to that in Louisville (entries are 10,000 times log of ratio) for 1937 for (January, ..., December) by day of week: Monday (88, 145, 148, 156, 133, 181, 208, 201, 203, 122, 186, 166). Tuesday (108, 176, 138, 171, 145, 156, 207, 231, 208, 163, 225, 194). Wednesday (121, 191, 151, 141, 132, 199, 239, 252, 196, 196, 187, 174). Thursday (104, 168, 169, 140, 144, 165, 245, 230, 156, 155, 174, 148). Friday (103, 185, 133, 170, 113, 185, 214, 260, 248, 185, 204, 157). Make an appropriate two-way analysis, with plots.

18. T. H. Meltzer (1962. Round-robin testing of the lead peroxide ignition method of assaying polyethylene oxidizability. Materials Research and Standards 2: 729-733) reported the ignition times (in hundredths of a minute) found for the same 5 test resins (1,2,3,5,6) by various laboratories. The results (his table II on page 732) for the first 6 laboratories were: Dupont (31, 34, 38, 43, 35). Electric Storage Battery (40, 40, 47, 49, 62). Eastman Chemical (32, 36, 43, 44, 38). Imperial (38, 37, 43, 49, 41). Koppers (22, 24, 30, 38, 24). National Petrochemicals (56, 56, 62, 66, 61). Make a reasonable and resistant two-way analysis. Draw a two-way plot of the results. Make a plot of the coded residuals.

19. Meltzer (see above) also reported results from 6 other laboratories. Owens-Illinois (46, 46, 52, 57, 49). Phillips (34, 34, 37, 42, 36). Picatinny (29, 31, 33, 34, 30). Spencer (39, 39, 40, 48, 41). Union Carbide at Bound Brook (37, 37, 42, 46, 40). Union Carbide at South Charleston (24, 24, 27, 29, 26). Make a resistant two-way analysis of these results. Make a two-way plot. Make a plot of coded residuals.

20. Compare the results of the two analyses above. How are they similar? How different? Combine the two sets of residuals into a single 6 by 12 table and take 2 half-steps of median polish. How much has the analysis changed? What changes, if any, in the conclusions are called for? Make a two-way plot of the coded residuals.

(The four problems that follow can easily be shared among 3 or 4 friends. The last of the four is not routine.)

21. Dockray, Meadows, and Smith (1950. Cotton wast. Textile Research Journal 20: 684-718) report, based on the Bureau of the Census, the exports (in thousands of pounds) of cotton waste for the years (1940 C 1942 E 1944 G 1945 I 1948) as follows: Card strips (2,604 C 532 E 1,341 G 970 I 710). Comber wast (15,642 C 2,920 E 12,128 G 6,145 I 14,658). Other soft wast (21,160 C 6,753 E 5,961 G 10,832 I 21,486). Hard waste (14,193 C 5,902 E 5,298 G 8,940 I 6,612). Make a reasonable and resistant two-way analysis. Make the associated two-way plot of the fit. Make the corresponding plot of coded residuals.

22. Dockray, Meadows, and Smith (see above) also give corresponding data on imports, namely: Card strips (3,180 C 424 E 60 G 4648 I 2774). Comber waste (177, none, none, 80, 109). Other soft waste (24,869 C 5705 E 747 G 106,098 I 57,334). Hard waste (633 C 1225 E 4,373 G 7,057 I 1,610). Make the same analyses and plots.

23. From the data of the last two problems calculate the net export (exports MINUS imports) of each type of waste in each year. Make the same analyses and plots.

24. Compare and combine the results of the last three problems. Discuss the results. Is there any way to make a single two-way table using all the results?